

Dual Clustering Detection Model of Power Theft Based on User Behavior Analysis

Dongdong Zhang, *Senior Member, IEEE*, Feng Jiang, Hui Hwang Goh, *Fellow, CSEE*, Tianhao Liu, *Member, IEEE*, Hui Liu, *Senior Member, IEEE*, and Thomas Wu, *Senior Member, IEEE*

Abstract—Under the background of big data, a series of power theft behaviors of residents not only cause significant economic losses but also bring incalculable security risks. According to the unique characteristics of residential users, a dual clustering detection model of power theft is proposed in this paper based on analysis of users' power consumption behavior. First, by establishing a preliminary diagnosis model of power theft based on improved K-means clustering, shortcomings of the original K-means algorithm, such as determination of the K value and initial cluster center, low calculation efficiency, and vulnerability to noise, are improved. Second, based on power consumption characteristics of residential users, their power consumption behavior patterns are classified, 11 types of residential power consumption patterns are summarized, and the final set of power thieves is determined by curve similarity analysis. Simulation results show the proposed model can significantly improve detection rate, false detection rate, and other evaluation indicators for the dataset of residential users at the same time, which provides a reference for power theft detection in power supply companies.

Index Terms—Anomaly detection, curve similarity comparison, K-means clustering, power consumption behavior.

I. INTRODUCTION

THE electrical energy lost by power companies during the process of production, transportation, and distribution can be divided into technical losses (TLs) and non-technical losses (NTLs) [1]. NTLs account for approximately 10%~40% of total power loss, and the main types include a series of power theft behaviors at the distribution side, meter failure, improper operation, and so on [2], [3], especially losses caused by power theft.

Based on incomplete statistics, economic losses caused by abnormal power consumption in China is approximately 10 billion yuan every year. Moreover, electricity theft in the United States reduces yearly revenue of power companies by \$6 billion. Currently, India has the most serious power theft in the world, with resulting economic costs as high

as US \$16.2 billion every year [4]–[6]. In recent years, several large-scale electricity thefts have emerged worldwide through bitcoin mining [7]. An example is the 2009 bitcoin mining case in Zhenjiang, Jiangsu Province, which involved tens of millions RMB of electricity theft. Abnormal power consumption behaviors significantly threaten safe and stable operation of the power grid [8]. According to data statistics, 40% of fire accidents and electric shock incidents in 2021 were a result of electricity thefts. Meanwhile, some power consumption enterprises improve their market competitiveness by stealing electricity, which creates bad enterprise production atmosphere [9].

Traditional means of electricity theft prevention is mainly through regular investigations by professionals and installation of monitoring or alarm instruments systems in the meter box. However, these methods not only waste considerable human and material resources but also have a very low efficiency [10], [11]. Today, with rapid popularization of smart meters, tens of thousands of smart meters have collected a large amount of power consumption data, weather data, user information, and economic conditions have been included in the data analysis scope [12]–[14]. Meanwhile, power thieves have devised elaborate, intelligent, and industrialized means of stealing electricity. Reference [15] described loopholes of smart meters and proposed defense strategies under different pricing schemes. However, useful information needs to be further extracted from a large amount of data, and data mining methods for detecting abnormal power consumption can roughly be divided into classification, regression, and clustering. Classification and regression are supervised learning techniques, whereas clustering is an unsupervised learning technique [16]–[18].

The key to adopting classification method for power consumption anomaly detection is to train a large amount of labeled data, extract hidden features, and select the appropriate classifier [19]. Reference [20] used several machine learning algorithms to detect abnormal power consumption of Tunisian customers using a large dataset comprising their invoices. They found efficiency of the integrated classifier was better than other algorithms and achieved good performance on the unbalanced dataset. Reference [21] used XGBoost integrated algorithm to detect six specific attacks and compared the result with eight common machine learning algorithms. Results showed the integrated algorithm had higher detection rate and lower false detection rate. With rapid development of deep learning, neural networks have been widely used in classification problems [22], [23]. However, most existing power

Manuscript received November 12, 2021; revised March 25, 2022; accepted May 10, 2022. Date of online publication June 27, 2023; date of current version August 17, 2024. This work was supported by National Natural Science Foundation of China (52567005).

D. D. Zhang, F. Jiang, H. H. Goh, H. Liu and T. Wu are with the School of Electrical Engineering, Guangxi University, Nanning 530004, China.

T. H. Liu (corresponding author, email: thliu@eee.hku.hk) is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China.

DOI: 10.17775/CSEEJPES.2021.08430

theft detection methods are based on one-dimensional power consumption data; hence, they have low detection accuracy and cannot capture the power consumption cycle. To tackle this problem, Reference [24] proposed a power theft detection method based on wide area and convolution extended network, and experimental results revealed the method had better performance in power theft detection of two-dimensional power consumption data than other methods. To compensate for vulnerability of machine learning algorithms, Reference [25] simulated a hidden power stealing strategy, determined consumption deviation at the family and interpersonal levels, and developed a detection model based on deep learning.

The core idea of anomaly detection using regression algorithm is load forecasting, which detects power theft based on deviation between actual power consumption and forecast. Reference [26] performed load forecasting based on the weighted average of historical data and observed the error between actual and predicted values to measure the degree of abnormal power consumption. Reference [27] used five daily load characteristics instead of a high-dimensional daily load curve to characterize the power consumption mode. They compared the reconstruction error of each sample with its adjacent samples and calculated the local outlier score to represent the user's anomaly degree. Reference [28] used the revised linear regression model to capture the physical relationship between power consumption and voltage on the secondary side of the distribution, and found that abnormal users and normal users would lead to negative residuals and positive residuals of the distribution side regression, respectively.

The idea of using clustering method to detect abnormal power consumption is to roughly classify similar ethnic groups in power consumption data into a cluster. Outliers that are remote from center and high-density points can then be regarded as suspected abnormal power users. Reference [29] combined the correlation analysis of line loss and clustering by fast search and find of density peaks for daily load curve and used receiver operating characteristic (ROC) curve and mean average precision as the evaluation indexes of the model. Reference [30] proposed two unsupervised anomaly detection methods based on rain flow. One method uses the connectivity outlier factor algorithm to solve the point anomaly problem, whereas the other method uses the nearest neighbor anomaly factor algorithm to solve the set anomaly problem. Refer-

ences [31], [32] proposed an NTL model using optimal path forest unsupervised learning algorithm and Gustafson-Kessel's fuzzy clustering algorithm. Results showed the corresponding model achieved good results. Reference [33] constructed a semi-automatic feature extraction method based on clustering, which uses short-term pattern to extract short-term power consumption rules for abnormal power consumption detection [34].

Although there has been a lot of research in this field, there are still many problems to be solved. In this paper, we propose a dual-cluster power theft detection model based on user power consumption behavior analysis, as shown in Fig. 1. The specific work and contributions are summarized as follows:

1) To solve the problem in which traditional power theft detection models cannot simultaneously determine detection and false detection rates. First, in the clustering process, the distance threshold is set at a relatively low value to ensure the model has a high detection rate. Subsequently, false detection rate of the model is reduced by analyzing users' power consumption behaviors.

2) To improve detection rate and calculation efficiency of the preliminary detection model of electricity theft using K-means clustering. In this paper, the double criteria of elbow and contour coefficient methods are used to determine the K value, K-means++ algorithm is used to select the initial clustering center, Elkan K-means is used to improve computational efficiency of the algorithm, and K-medoids is used to reduce impact of noise points on clustering. Compared to other clustering algorithms, effectiveness of the improved K-means algorithm is verified.

3) To solve the problem of incomplete extraction and poor interpretability of traditional users' power consumption features. In this paper, power consumption behaviors of users are classified in detail according to characteristics of residential users' power consumption dataset. Additionally, their power consumption characteristics are fully grasped, feature extraction accuracy is improved, and feature extraction effectiveness is verified by comparison with traditional machine learning methods.

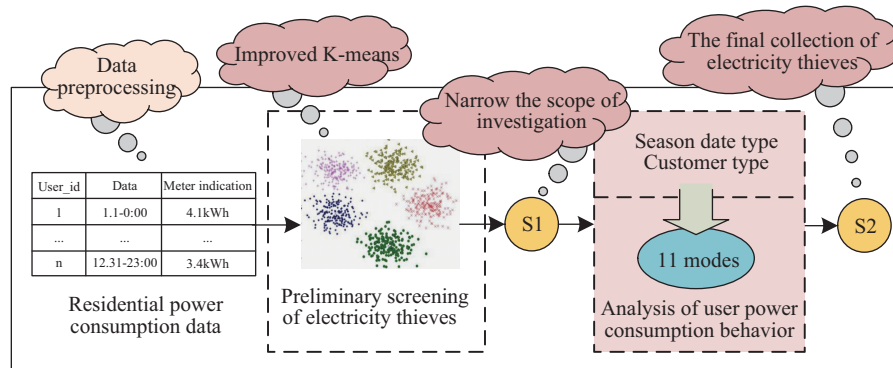


Fig. 1. Frame diagram of full-text model.

II. PRELIMINARY DIAGNOSIS OF POWER THIEVES BASED ON IMPROVED K-MEANS ALGORITHM

To narrow the scope of screening electricity thieves, a preliminary diagnosis method for detecting electricity thieves is proposed in this section based on improved K-means clustering algorithm. In cluster analysis, an attempt is made to find the relationship between data objects, and data are grouped. The more similar the data in the group, the greater the difference between the groups, the better the clustering effect. As the most widely used clustering algorithm, K-means has the advantages of simple principle, convenient implementation, and fast convergence. However, its K value is difficult to determine, and the initial clustering center is difficult to determine. In addition, it is vulnerable to noise and its computing efficiency for large data is low. Using the load characteristics of residential users, the K-means clustering algorithm is improved in this section to a certain extent to enhance its effect in the preliminary screening of power theft.

A. Problems of Traditional K-means Algorithm and Improvement Method

The idea of K-means algorithm is very simple. For a given sample set, the sample set is divided into K clusters according to the distance between samples is too small, so the points in the cluster are connected as closely as possible, and distance between clusters is as large as possible [35], [36].

K-means has advantages of simple principle, convenient implementation, and strong interpretability. However, its several disadvantages cannot be ignored. First, K value is usually determined randomly based on experience; however, K value is generally difficult to estimate, and effects of different K values are also very different. Second, the initial clustering center has

significant impact on the clustering effect. An inappropriate initial clustering center will not only reduce convergence speed of the algorithm but also make the algorithm fall into local optimum. Third, traditional Euclidean distance significantly reduces calculation efficiency owing to the large sample of residential power consumption data. Finally, the algorithm is vulnerable to noise. Hence, the K-means algorithm is improved in Section II-B from the following aspects:

- 1) The double criteria of elbow [36] and contour coefficient methods [37] are utilized to select the K value;
- 2) K-means++ is used to improve selection of the initial clustering center;
- 3) Elkan K-means is used to improve efficiency of the algorithm;
- 4) K-medoids are used to reduce the interference of noise in the clustering process.

B. Preliminary Diagnosis Model of Power Theft Based on Improved K-means Algorithm

To preliminarily screen electricity thieves, a preliminary diagnosis model of electricity theft is established based on improved K-means algorithm. Figure 2 shows the main architecture of the model. First, original data were pre-processed and K value was determined using the double criteria, as shown in Fig. 2(b). Second, an algorithm flowchart for determining the initial clustering center was designed, as shown in Fig. 2(c). Finally, after the K clusters were obtained, outlier set S1 was regarded as the undetermined set of electricity thieves.

As shown in Fig. 2(b), the core idea of the elbow method is that as cluster number K increases, core index SSE will gradually decrease, as given by (1). When the value of SSE

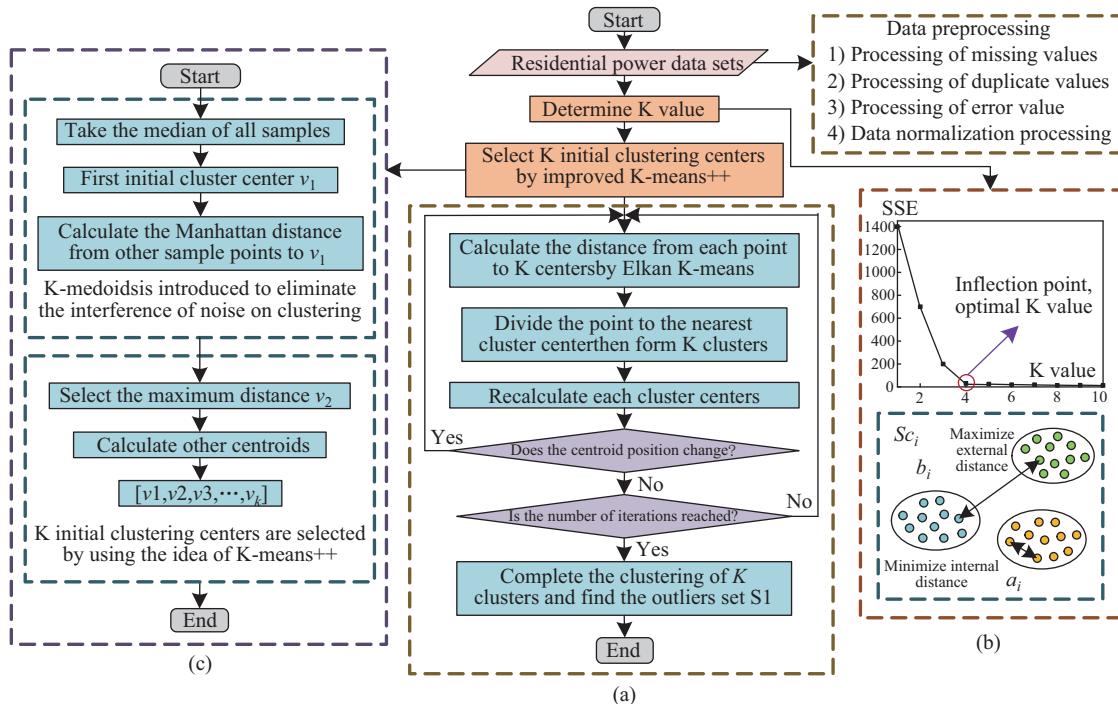


Fig. 2. Flow chart of improved K-means clustering algorithm. (a) Outlier detection module. (b) Double criterion of K value. (c) Selection of initial clustering center.

is close to the elbow, the best K value is obtained.

$$SSE = \sum_{i=1}^K \sum_{P \in C_i} |P - m_i|^2 \quad (1)$$

where C_i is the i^{th} cluster, P is the sample point in C_i , m_i is the centroid of C_i , and SSE is the clustering error of all samples, representing the quality of the clustering effect.

In the contour coefficient method given by (2), each point i is the sample in the known cluster data, and a_i represents average distance from i to its own cluster, whereas b_i represents minimum distance of all samples of other populations, following the principle of “high cohesion and low coupling”:

$$Sc_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (2)$$

where Sc_i is the contour coefficient. When $b_i \gg a_i$, Sc_i tends to 1, and the clustering effect is better. When $b_i \ll a_i$, Sc_i tends to -1 , the clustering effect is worse.

However, both elbow and contour coefficient methods have their disadvantages. The inflection point of the elbow method is difficult to determine. Currently, most methods for determining the inflection point are visual-based; however, the inflection point is determined based on constant slope. Moreover, the optimal number of clusters selected by the contour coefficient method may lead to high SSE. Therefore, the elbow and contour coefficient methods are combined in this paper to select the best K value.

As shown in Fig. 2(c), to solve the problem in which K-means algorithm randomly selects the initial cluster center, resulting in a significant difference between clustering results and actual data distribution, K-means++ algorithm was used to improve selection of the initial cluster center. For residential power datasets, division of clusters will be affected by noise to a certain extent. Thus, K-medoids algorithm was introduced into the K-means++ algorithm to reduce impact of outliers on the division results. At the same time, in the traditional K-means algorithm, distance from sample points to all centroids needs to be calculated in each iteration, which significantly reduces calculation efficiency. To reduce calculation time of the model, Elkan K-means algorithm was introduced, which uses the idea the sum of both sides of the triangle is greater than the third side and the sum of both sides is less than the third side, to reduce the amount of distance calculation. Fig. 2(c) shows the flowchart of the improved algorithm, and specific flow of the algorithm is as follows:

- 1) The original data is preprocessed, including missing values, duplicate values, error values, and data standardization.
- 2) The double criteria of the elbow and contour coefficient methods are used to determine the K value.
- 3) The first initial cluster center v_1 is obtained by calculating the median of all samples.
- 4) The Manhattan distance from all sample points to the first initial cluster center v_1 is calculated, maximum distance L_i is obtained, and the sample point with the largest distance is selected as the second initial cluster center v_2 .
- 5) The third initial cluster center v_3 is calculated.
- 6) Steps 4) and 5) are repeated to find K initial clustering centers $[v_1, v_2, v_3, \dots, v_k]$.

7) Distance from each point to the K centroids is calculated and K clusters are formed using Elkan K-means.

8) Iterate repeatedly until the position of the cluster center is no longer offset or the maximum number of iterations is reached.

9) The threshold is set to determine the outlier set S1.

After clustering, a threshold needs to be set to determine the set of outliers. Because this model preliminarily determines power thieves in overall residential power consumption data, this threshold needs to be slightly smaller and more suspicious objects need to be identified as power thieves to reduce missed detection rate. The threshold problem will be explained in Section IV. After the preliminary diagnosis model processed residential power consumption data, an outlier set S1 was obtained, which was used as input of the user power consumption behavior analysis model to reduce false detection rate of the overall model.

III. CLASSIFICATION OF USER POWER CONSUMPTION BEHAVIOR PATTERNS

In view of the collected S1 of pending electric larceny in the previous section, larceny needs to be further detected to reduce false detection rate of the model. In this section, we establish a power theft detection model based on user power consumption analysis. First, through exploration and analysis of residential power consumption data, we determine factors that can mostly affect residential power consumption behavior mode. Second, we establish a classification model of the residential power consumption mode and determine residential power consumption modes in 11. Finally, abnormal points of different power consumption modes are determined by curve similarity comparison to obtain suspect set S2, where $S2 \subseteq S1$.

A. Analysis of Factors Affecting Residential Power Consumption Behavior

Living habits of residents often determine their power consumption behavior [38]. Owing to alternating action of various factors, residents' power consumption behavior has strong randomness. Factors affecting residents' electricity consumption behavior are generally divided into external factors, such as temperature, holidays, and emergencies, and internal factors such as family member structure and income. Some of these factors directly form a specific power consumption mode. For example, extreme temperature may directly affect air conditioning load. In contrast, small probability events such as emergencies have a slight impact on the overall power consumption mode and can be ignored.

Seasonal changes produce extreme temperatures, which can directly affect use of some loads like air conditioning. To further analyze influence of season on power consumption mode, daily load curve of different types of families in different seasons on weekdays and thermodynamic diagram of air conditioning load are shown in Fig. 3.

As shown in Fig. 3(a)–(c), users of different types of families, overall power consumption in summer and winter was much greater than in spring and autumn, and it was more

obvious in three peaks. Heat maps are shown in Fig. 3(d)–(f), indicating the air conditioning load of various types of users is mainly concentrated in summer and winter. Therefore, the seasonal factor is an important basis for dividing power consumption mode for residential users, and spring and autumn seasons can be regarded as one category, while summer and winter seasons can be regarded as another category to simplify classification standard.

For a residential area, regardless of vacant houses, family types of users are roughly divided as follows: office worker families, elderly families, and mixed families composed of workers and the elderly. For convenience of description, office worker families, elderly families, and mixed families are numbered as class I, class II, and class III families, respectively. Figure 4 shows power load of different types of households.

As shown in Fig. 4, the power consumption curve of users in residential areas can roughly be divided into three categories. Figure 4(a) shows power consumption of office

worker families. Power consumption behavior of such users was very consistent, and the power consumption peak occurred at 7:00–9:00 a.m., 12:00–14:00 p.m. and after 18:00 p.m. Figure 4(b) shows power consumption behavior mode of elderly families. The overall power consumption of users tended to be stable, and power consumption in peak hours was significantly less than office worker families. Figure 4(c) shows power consumption of mixed families. Overall power consumption of these users was significantly greater than of office worker and elderly families and combined characteristics of three peak hours of office worker families and non-peak hours of elderly families.

For office worker families, there were significant differences in power consumption behavior on weekdays and holidays. Although there were many kinds of holidays, power consumption behavior of users on holidays was easy to analyze. Figure 5 shows three typical power consumption behaviors of users on holidays.

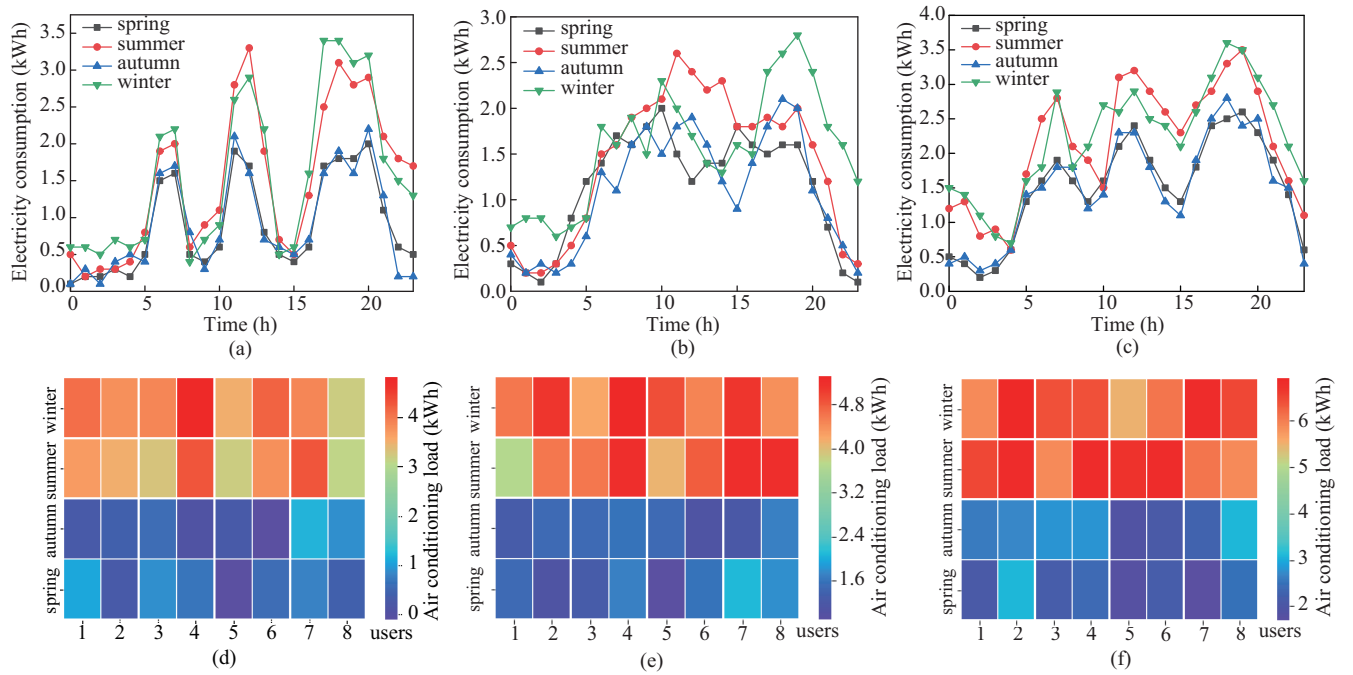


Fig. 3. Four seasons load curve and air conditioning load heat map of different types of users. (a) Power load of officers in four seasons. (b) Power load of elders in four seasons. (c) Power load of mixed family in four seasons. (d) Heat map of air conditioning load of office workers' families. (e) Heat map of air conditioning load of elders' families. (f) Heat map of air conditioning load of mixed family families.

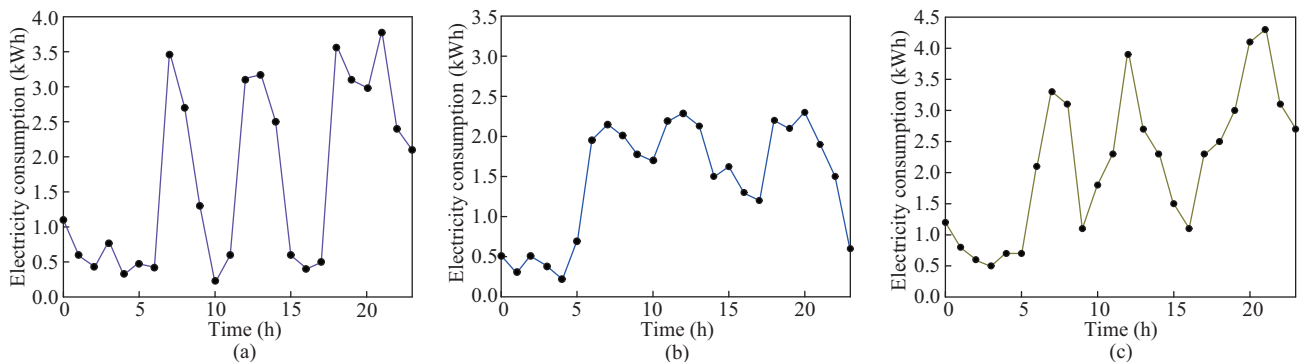


Fig. 4. Comparison of daily power load of different household use. (a) Daily load curve of class I users. (b) Daily load curve of class II users. (c) Daily load curve of class III users.

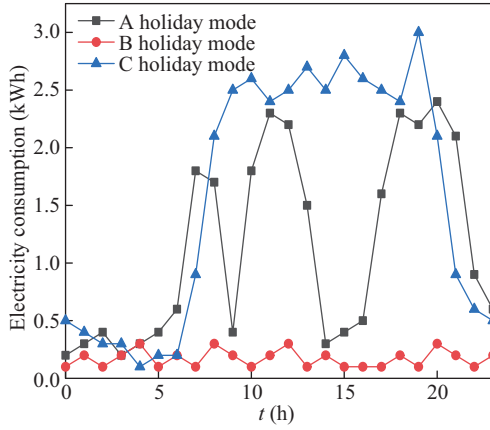


Fig. 5. Comparison of daily power load in different types of holidays.

From Fig. 5, there were roughly three kinds of power consumption behaviors during holidays. Class A curve represents power consumption during working days, and it had three peaks. It is concluded users on such holidays work normally and do not take leave. For class B curve, overall power consumption tended to be stable and below 0.5 kWh; hence, it is concluded user may travel with his family. For class C curve, power consumption peaks were from 8:00 a.m. to 21:00 p.m.; hence, it is concluded that users rest at home on such holidays.

B. Classification Model of User Behavior Patterns

Considering suspect set S1 obtained by preliminary clustering in the previous section, a power theft detection model is developed in this section based on user power consumption behavior analysis. According to the season, user type, and date type in the previous section, user's power consumption data were classified and filtered, and 11 power consumption modes were obtained. Afterward, users that were not captured within the classification results were classified as abnormal point set S2, which had a greater probability of power theft. Figure 6 shows the model architecture.

After classifying dataset S1 according to season, user type, and working day, and holiday type, the classification network and classification category were obtained, as shown in Fig. 6. Among them, A and B represent spring and autumn and summer and winter seasons, respectively, I, II and III represent three different family types, and A, B and C represent three different power consumption modes on holidays. As shown in Fig. 6, 18 power consumption modes were randomly arranged and combined with various factors. However, some of them were repeated, such as the elderly family does not work; hence, they can be classified into one category. Second, regardless of season or family, in the case of collective travel on holidays, power consumption behavior patterns of users were very similar; hence, they can also be classified into one category. Through the above analysis, power consumption behavior patterns in residential areas were divided into 11 categories, and points that were not captured within the above classification results were classified as abnormal point set S2.

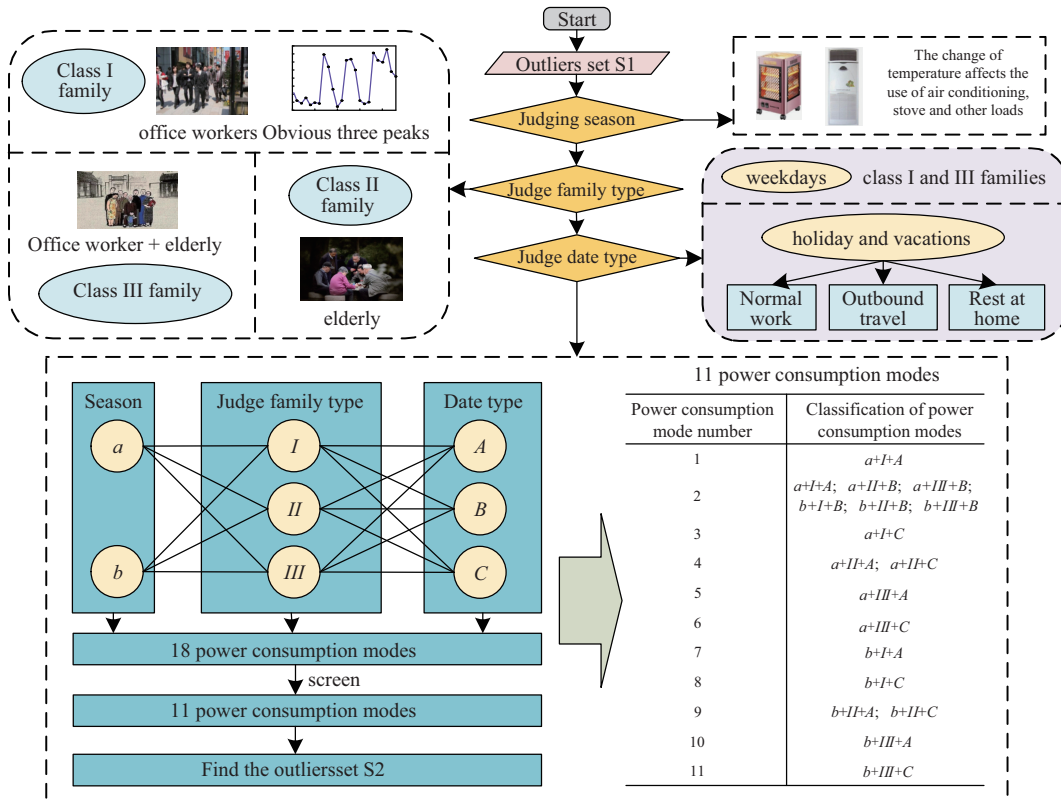


Fig. 6. Abnormal power consumption detection model based on user power consumption behavior analysis.

C. Determination of the Electricity Thief Set S2

Generally, to measure similarity of two curves, the common method is to compare distance and shape. However, in this study, the shapes of two users in the same power consumption mode do not necessarily have high similarity. Therefore, curve similarity in this model was considered from the perspective of distance. General Euclidean distance [39] holds that for sequences with the same length, distance between each two points is calculated and then summed. The smaller the distance, the higher the similarity. However, the goal of this paper is to detect power thieves. The ultimate goal of power thieves is to reduce their own power consumption cost, it is assumed at time t , the actual power consumption of user is a_t , and after being tampered by the power stealing user, the meter display becomes b_t . When the corresponding unit price is p_t , considering period T , the purpose of the power thief can be expressed by (3):

$$\sum_{t \in T} p_t b_t \leq \sum_{t \in T} p_t a_t \quad (3)$$

It can be seen the Euclidean distance cannot be directly used to define distance. Distance can only be defined when actual power consumption is less than expected mode power consumption, the part where the actual curve is greater than the mode curve is defined as 0. Therefore, distance between curves is defined by (4):

$$l = \begin{cases} l_{\text{mode}} - l_{\text{acture}}, & l_{\text{mode}} > l_{\text{acture}} \\ 0, & l_{\text{mode}} \leq l_{\text{acture}} \end{cases} \quad (4)$$

where l_{mode} represents the curve obtained after classification of power consumption mode behavior and l_{acture} represents the user's actual power consumption curve.

When distance l exceeds a certain length unit m , the sequence is placed into abnormal sequence set S2'. In addition, for a sequence, when actual power consumption b' is less than a certain proportion b of power consumption of the applied power mode, it should also be placed in abnormal sequence set S2'. Finally, if the abnormal sequence exceeds a certain number of days n within 30 days, it is considered to be the final power thief and is put into set S2. Figure 7 shows the process framework.

D. Model Evaluation Index

Owing to imbalance of data, result of the classifier is more likely to be biased toward most categories. For example, assume there are two power theft users among 100 users. If remaining users are detected as normal, accuracy rate will be 98%, which is meaningless. Therefore, recall rate, error rate, F1-score, ROC curve, and AUC were used as evaluation indexes, and calculation of these indexes relies on the confusion matrix [40], which is expressed in Table I:

TABLE I
CONFUSION MATRIX

Description	Forecast normal	Forecast anomaly
Actually normal	TP	FN
Actually anomaly	FP	TN

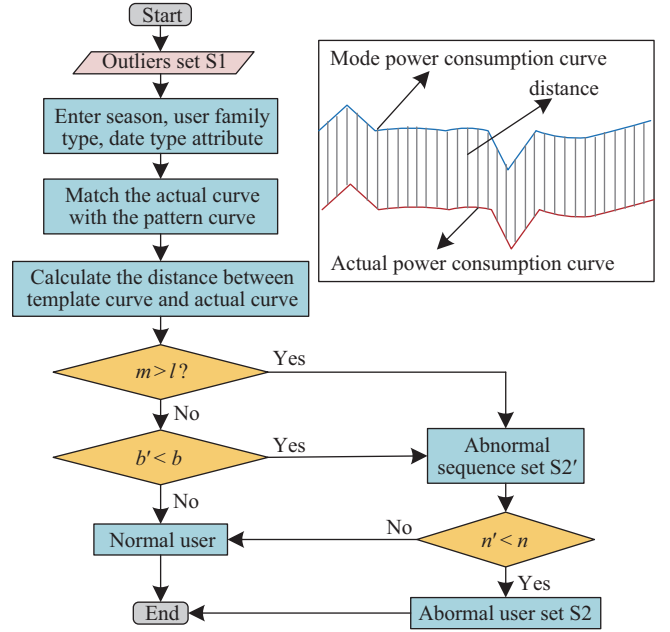


Fig. 7. Determination of abnormal power consumption set S2.

Calculation formulas for various indexes were obtained from Table I, and they are given by (5)–(9):

$$Pre = \frac{TP}{TP + FP} \quad (5)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

$$Rec = \frac{TP}{TP + FN} \quad (7)$$

$$ERR = \frac{FP}{TN + FP} \quad (8)$$

$$F_1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} \quad (9)$$

where Pre represents the proportion of the number of correct targets detected to the number of detected targets. However, all categories should be considered when calculating accuracy. Therefore, accuracy of the model is measured by the ACC index. Rec represents the proportion of all normal users who are identified as normal users. ERR indicates the proportion of abnormal users divided into normal users. F_1 considers both precision and recall, and weight is equal.

IV. SIMULATION EXAMPLE ANALYSIS

A. Data Sets and Data Preprocessing

In this study, the power consumption dataset of residential users in a community was used as the research object. Except for vacant houses, the dataset contains 3234 data samples, including 2893 normal users and 341 abnormal users. Data collection interval was once every 60 minutes. Original data were mixed with a lot of noise to improve data quality, and they were preprocessed as follows: For duplicate data, deletion strategy was directly adopted. Missing data were generally composed of many zero values and individual missing values. Continuous occurrence of multiple missing values with zero has little significance for data filling, and deletion method

was directly adopted. For a single missing value, power consumption data of five days before and after the missing value are considered to form a group, which was filled by Lagrange interpolation method [41].

B. Determination of K Value in K -means Algorithm

To solve the problem of random selection of the K value, we analyzed SSE and S_{ci} values under different K values. Figure 8 shows the elbow and contour coefficient diagrams under different K values.

As can be seen from Fig. 8(a), the best K value obtained by the elbow method was between 4 and 5, but the specific one is more suitable as the K value cannot be judged subjectively. In Fig. 8(b), among K values obtained by the contour coefficient method, when K values were 2 and 4, the contour coefficient value was large. However, when K was 2, it was obviously inconsistent with the selection of elbow method. Table II presents calculation efficiency of K from 1 to 10.

As can be seen from Table II, in the range of 1–10, calculation time gradually increased with an increase in K ; however, increase speed was fast at the beginning. When K value exceeded 5, increase speed of calculation time slowed down accordingly. Therefore, by combining the two methods, the best K value was 4.

C. Determination of Clustering Threshold and Outliers set $S1$

When preliminarily determining power thieves, it is necessary to set a threshold, which is the relative distance from the cluster center. When the relative distance is greater than this value, it is considered as an abnormal point, and when the relative distance is less than this point, it is considered as a normal user. Figure 9 shows set distribution of abnormal power consumption under different thresholds.

As shown in Fig. 9, the number of abnormal points corresponding to different thresholds was different. The smaller the threshold set, the more abnormal points exist, accordingly, false detection rate will increase, and missed detection rate will decrease. When preliminarily determining suspects, suspects should be defined as abnormal points as much as possible, that is, a certain false detection rate can be sacrificed to reduce missed detection rate, and false detection rate can be reduced through the subsequent model. Indicators corresponding to different thresholds are presented in Table III. It can be seen from the table when the threshold was 2, the final missed detection rate decreased to 0.0071. Therefore, the clustering threshold was set to 2 in this study to obtain suspect set $S1$.

D. Determination of Electricity Thief Set $S2$

To further determine the set of electricity thieves, user power consumption data of each power consumption type

TABLE II
INFLUENCE OF DIFFERENT K VALUES ON CALCULATION TIME

K value	Calculation time (s)	K value	Calculation time (s)
1	527	6	1369
2	1033	7	1387
3	1167	8	1396
4	1237	9	1419
5	1322	10	1424

TABLE III
INDICATORS CORRESPONDING TO DIFFERENT THRESHOLDS

Threshold	Number of outliers	Number of normal points	Missed detection rate	False detection rate
2	174	3060	0.0071	0.1248
2.5	65	3169	0.0825	0.0447
3	29	3205	0.1543	0.0128
3.5	8	3226	0.2863	0.0035

was averaged, and 11 power consumption mode curves were obtained, as shown in Fig. 10.

From the figure, the 11 power consumption modes were roughly divided into the following types. First, mode 2 was special and its overall power consumption was below 0.4 kWh, indicating users are not at home. Second, when divided by season, modes 1–6 (excluding 2) were in spring and autumn, and 7–11 were the user's summer and autumn power load curves. From the order of magnitude, power consumption of modes 7–11 were obviously greater than those of 1–6. When divided by date type, office worker families had three obvious power consumption peaks during working days, which were marked with red circles in Fig. 10. For date type of rest at home and elderly families, their power consumption curve had a continuous peak in the middle time period, and the corresponding curve was marked with purple lines. Finally, there were also differences in the similar curve as power consumption of mixed families was greater than office worker and elderly families.

In the process of determining final power thief set $S2$, we need to compare and analyze curve similarity between user's real power consumption data curve and the above corresponding power consumption mode. In this process, some parameters need to be determined: (1) Distance between curves l . (2) When distance exceeds m length units, it is recognized as an abnormal sequence. (3) When determining the sequence of power theft behavior, actual power consumption accounts for the percentage of power consumption in corresponding mode b . (4) When the abnormal sequence of a user exceeds a certain number of days n in a month, the user is considered a power thief. The above four parameters restrict one another to a certain extent. For example, between l and m , if the value of l is set larger, m can be set smaller accordingly. To reduce complexity of the algorithm, $l = 1$ and $n = 10$ were set according to data characteristics, and m and b were determined under this premise to achieve parameter balance.

From Table IV, as the value of m increased, fewer sequences were identified as abnormal sequences, missed detection rate

TABLE IV
COMPARISON OF UNDETERMINED EXPERIMENTAL PARAMETERS

m	Missed detection rate	False detection rate	b	Missed detection rate	False detection rate
8	0.0217	0.1698	0.1	0.5217	0.0032
9	0.0343	0.1324	0.2	0.4683	0.0149
10	0.0416	0.1047	0.3	0.3672	0.0423
11	0.0471	0.0738	0.4	0.2083	0.0631
12	0.0504	0.0461	0.5	0.1369	0.0877
13	0.0833	0.0431	0.6	0.0817	0.0936
14	0.1382	0.0318	0.7	0.0264	0.1295
15	0.1761	0.0211	0.8	0.0018	0.1618

increased gradually, and corresponding false detection rate decreases. When the value of m was 12, missed detection and false detection rates were within an acceptable optimal range. Parameter b was just opposite to m when b gradually increased, indicating more sequences were identified as power thieves, missed detection rate decreased, and false detection rate increased. To find a balance, b as taken as 0.6 for model construction. To sum up, on the premise distance is set to 1, when sequence length beyond 1 reached 12 or overall power consumption was only 60% of normal power consumption, the sequence of that day was identified as an abnormal sequence and placed in set S2'. When the number of abnormal sequences of a user reached 10 in a month, it was considered to be the final power thief and was then placed in set S2.

E. Evaluation of Model Effect

There were 3234 data samples in the resident user dataset, including 2893 normal users and 341 abnormal users, leading to a serious imbalance problem. Common methods for solving the problem of data imbalance are oversampling and under

sampling. Under sampling aims to balance distribution by reducing most class samples. Because it deletes data observations from original data, it has the risk of discarding useful information and is not suitable for use in the power thief detection model. Oversampling and under sampling are the opposite of each other, and they balance distribution by increasing the number of minority samples. Smote algorithm [42] is an improved algorithm for random oversampling. Its basic idea is to analyze minority samples and synthesize new samples according to minority samples to add to the dataset. It avoids the overfitting problem caused by random oversampling and adds a small number of samples by copying samples.

In this study, Smote algorithm was used to process unbalanced data, and the number of abnormal samples was increased to 2343. After samples passed the model test, the number of abnormal samples in the prediction results was 2853, and confusion matrix results are presented in Table V:

Evaluation indexes of the model were calculated from the confusion matrix, and results are presented in Table VI.

First, to verify effectiveness of the improved clustering algorithm proposed in this paper, we compared performance

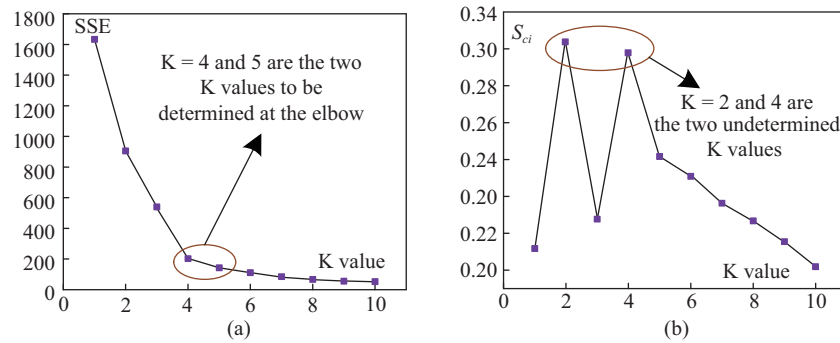


Fig. 8. Determination of the K value by the elbow and silhouette coefficient methods. (a) Elbow method. (b) Silhouette Coefficient method.

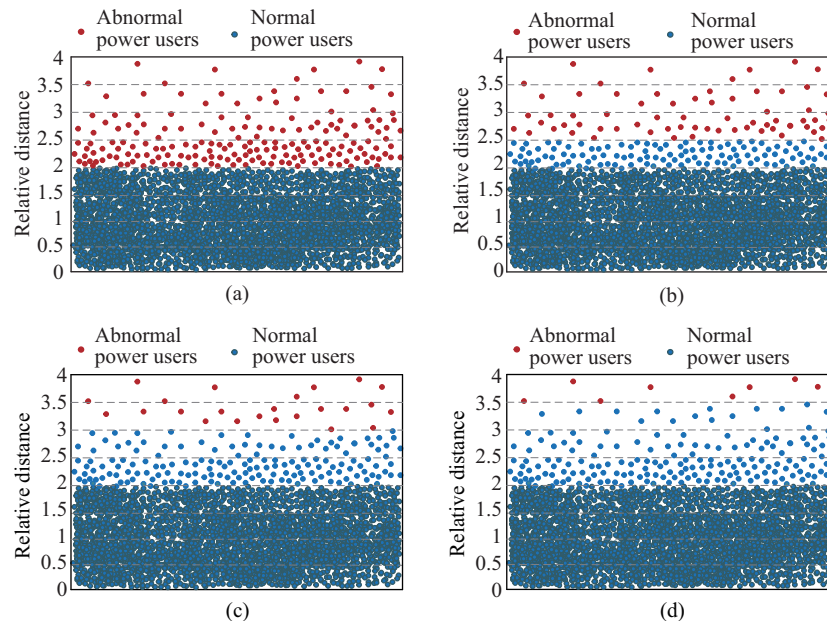


Fig. 9. Distribution of outliers under different thresholds. (a) Power consumption data distribution when the relative distance is 2. (b) Power consumption data distribution when the relative distance is 2.5. (c) Power consumption data distribution when the relative distance is 3. (d) Power consumption data distribution when the relative distance is 3.5.

TABLE V
CONFUSION MATRIX OF TEST RESULTS

Description	Forecast normal	Forecast anomaly
Actually normal	TP 2784	FN 109
Actual anomaly	FP 69	TN 2274

of the improved K-means algorithm with those of traditional K-means, K-means++, Elkan K-means and K-medoids algorithms in terms of operation efficiency, detection rate, and false detection rate. Figure 11 shows results.

As shown in Fig. 11, in terms of detection time, the improved K-means algorithm was tens of seconds slower than the Elkan K-means and exhibited good performance in

TABLE VI
VARIOUS EVALUATION INDEXES OF THE MODEL

Model evaluation index	Data results
<i>Acc</i>	0.96600
<i>Err</i>	0.02945
<i>Rec</i>	0.96232
<i>Pre</i>	0.97581
<i>F₁</i>	0.96902

accuracy, false detection rate, and F1-score.

To further verify performance of the proposed model, performances of classical machine learning, integrated learning, and clustering algorithms (without user power consumption

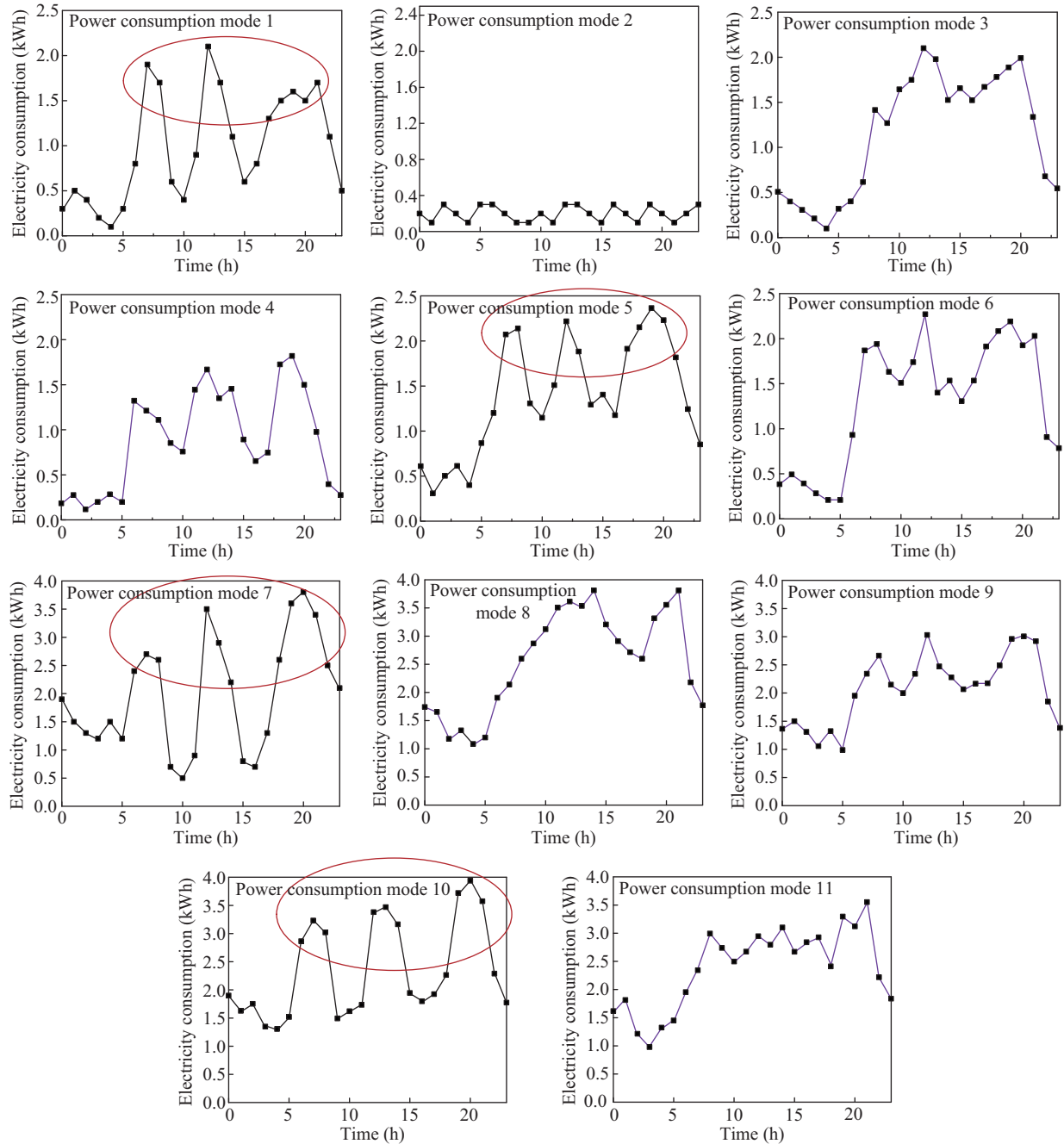


Fig. 10. Power consumption modes of 11 users.

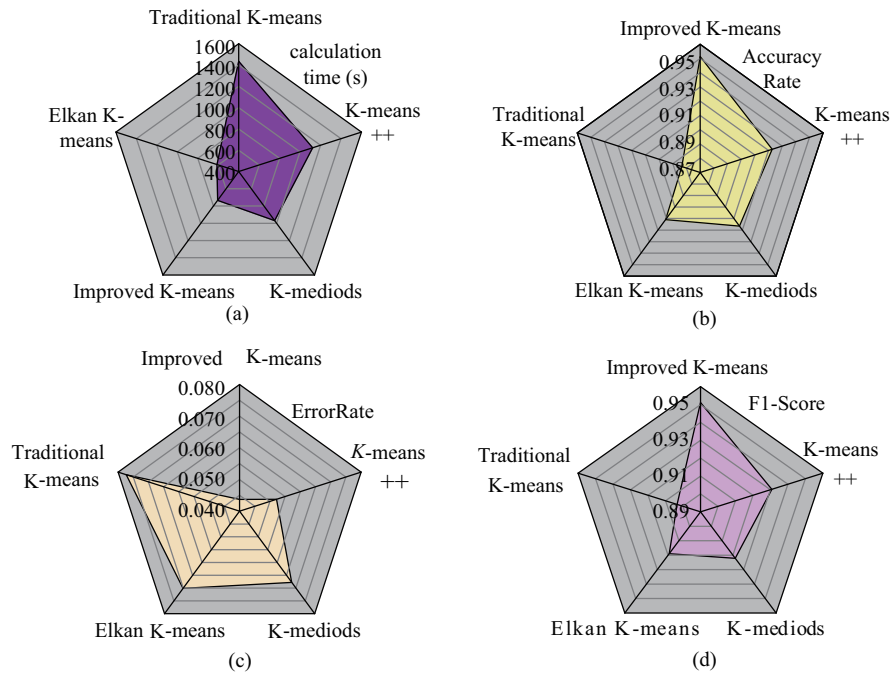


Fig. 11. Performance comparison of different K-means methods. (a) Comparison of calculation time of different K-means algorithms. (b) Comparison of Accuracy Rate of different K-means algorithms. (c) Comparison of Error Rate of different K-means algorithms. (d) Comparison of F1-Score of different K-means algorithms.

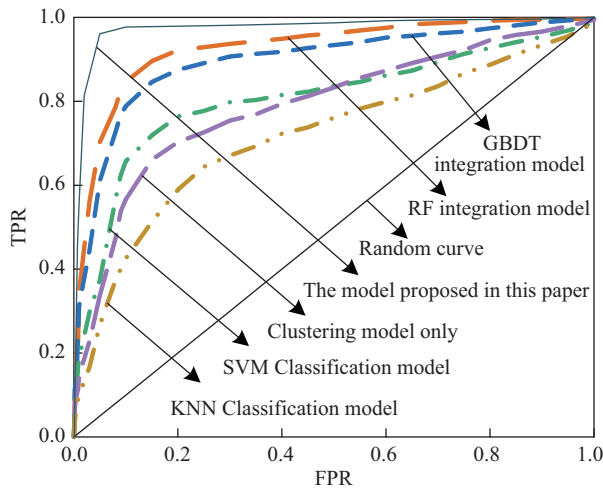


Fig. 12. ROC curve of model comparison.

behavior analysis) were compared based on the same user power consumption data mentioned earlier in this section. Steps of using machine learning algorithm to detect abnormal power consumption are as follows: First, feature extraction was performed on the preprocessed data. The feature extraction method was based on time series, mainly including a series of statistical feature indexes such as mean and variance, rising and falling trend indexes, Fourier transform frequency domain indexes, and so on. Second, machine learning classifiers such as KNN, SVM, or integrated learning algorithms such as RF were used as classifiers to classify normal and abnormal power users, and an ROC curve was obtained, as shown in Fig. 12.

It can be seen the AUC-ROC curve of the proposed model was significantly larger than of the above machine learning

classifiers and clustering algorithm, which shows the proposed model had better performance.

V. CONCLUSION

To investigate a series of electricity theft behaviors of residential users based on a large amount of data generated by smart meters, a dual-cluster electricity theft detection model based on user electricity consumption behavior analysis was proposed in this paper. First, based on original residential power consumption data, the model preliminarily locks the range of suspected power theft by improving the K-means outlier detection algorithm. Second, according to power consumption characteristics of residential users, power consumption behavior patterns of users were classified. Finally, the final set of power thieves was determined by curve similarity analysis. Through further analysis of results, the following conclusions were drawn:

1) Different types of datasets have their own characteristics. For different datasets, modeling should be performed after fully analyzing characteristics of datasets to achieve better results. After analyzing residential power consumption data, a model was proposed according to data characteristics, which had significant improvement compared to general data mining, feature extraction, and some common machine learning methods.

2) For the abnormal point detection model of power theft detection, both the missed and false detection rates were reduced by the proposed method. Through multi-layer detection, there were sufficient reasons and evidence to achieve the purpose of no false detection while ensuring that power thieves will not go unpunished.

3) In view of the difficulty of selecting appropriate K values in K-means clustering, the double criteria of elbow and contour coefficient methods can be used to find accurate K values. In addition, the proposed method can be used to determine the initial clustering center.

However, there are still some problems with this model. A series of power consumption behavior analysis of residential users in this paper cannot be used to represent power consumption behavior patterns in all regions. For different power consumption datasets, it is necessary to reanalyze datasets to construct different power consumption behavior patterns. Moreover, some parameters in this paper were not determined in the best way possible. For example, in Section IV-D, there is still room for improvement in the determination method of l , m , b , n , and so on. Therefore, construction of more adaptive models for different scenarios and development of diversified evaluation indicators will be hot research topics in the future.

REFERENCES

- [1] C. Li, X. Li, T. Li, Q. Meng, and W. Yu, "LMedS-based power regression: an optimal and automatic method of radiometric intercalibration for DMSP-OLS NTL imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2046–2057, Jan. 2021.
- [2] P. Laurinec and M. Lucká, "Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting," *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 413–445, Mar. 2019.
- [3] D. H. Yao, M. Wen, X. H. Liang, Z. P. Fu, K. Zhang, and B. J. Yang, "Energy theft detection with energy privacy preservation in the smart grid," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7659–7669, Oct. 2019.
- [4] H. Y. Qin, H. P. Zhou, and J. W. Cao, "Imbalanced learning algorithm based intelligent abnormal electricity consumption detection," *Neuro-computing*, vol. 402, pp. 112–123, Aug. 2020.
- [5] M. I. Ibrahim, M. Nabil, M. M. Fouda, M. M. E. A. Mahmoud, W. Alasmay, and F. Alsolami, "Efficient privacy-preserving electricity theft detection with dynamic billing and load monitoring for AMI networks," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1243–1258, Jan. 2021.
- [6] T. Ahmad, "Non-technical loss analysis and prevention using smart meters," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 573–589, May 2017.
- [7] E. Zaghloul, T. T. Li, M. W. Mutka, and J. Ren, "Bitcoin and blockchain: security and privacy," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10288–10313, Oct. 2020.
- [8] S. A. Salinas and P. Li, "Privacy-preserving energy theft detection in microgrids: a state estimation approach," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 883–894, Mar. 2016.
- [9] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019.
- [10] P. P. Biswas, H. Y. Cai, B. Zhou, B. B. Chen, D. Mashima, and V. W. Zheng, "Electricity theft pinpointing through correlation analysis of master and individual meter readings," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3031–3042, Jul. 2020.
- [11] K. D. Zheng, Q. X. Chen, Y. Wang, C. Q. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1809–1819, Mar. 2018.
- [12] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.
- [13] A. Takiddin, M. Ismail, M. Nabil, M. M. E. A. Mahmoud, and E. Serpedin, "Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4189–4198, Sep. 2021.
- [14] "Personalized federated learning for individual consumer load forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 1, pp. 326–330, Jan. 2023.
- [15] Y. Liu, T. Liu, H. Sun, K. H. Zhang, and P. F. Liu, "Hidden electricity theft by exploiting multiple-pricing scheme in smart grids," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2453–2468, Jan. 2020.
- [16] S. X. Fan, J. B. Guo, L. X. Li, G. Z. Wang, H. T. Xu, J. Yang, Z. N. Ning, "Framework and Key Technologies of Human-machine Hybrid-augmented Intelligence System for Large-scale Power Grid Dispatching and Control," *CSEE Journal of Power and Energy Systems*, vol. 10, no. 1, pp. 1–12, Jan. 2024.
- [17] J. I. Guerrero, I. Monedero, F. Biscarri, J. Biscarri, R. Millán, and C. León, C, "Non-technical losses reduction by improving the inspections accuracy in a power utility," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1209–1218, Mar. 2018.
- [18] K. Li, Y. Wang, N. Zhang, F. Wang and C. Huang, "Spatio-temporal granularity co-optimization based monthly electricity consumption forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 5, pp. 1980–1984, Sep. 2023.
- [19] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.
- [20] S. V. Oprea and A. Băra, "Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets," *Computers & Electrical Engineering*, vol. 94, pp. 107329, Sep. 2021.
- [21] Z. Z. Yan and H. Wen, "Electricity theft detection base on extreme gradient boosting in AMI," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 2504909, Jan. 2021.
- [22] M. Ismail, M. F. Shaaban, M. Naidu, and E. Serpedin, "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3428–3437, Jul. 2020.
- [23] H. B. Gooi, T. J. Wang and Y. Tang, "Edge Intelligence for Smart Grid: A Survey on Application Potentials," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 5, pp. 1623–1640, Sep. 2023.
- [24] Z. B. Zheng, Y. T. Yang, X. D. Niu, H. N. Dai, and Y. R. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.
- [25] L. Cui, L. Guo, L. X. Gao, B. R. Cai, Y. Y. Qu, Y. P. Zhou, and S. Yu, "A covert electricity-theft cyberattack against machine learning-based detection models," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7824–7833, Nov. 2022.
- [26] H. Janetzko, F. Stoffel, S. Mittelstädt, and D. A. Keim, "Anomaly detection for visual analytics of power consumption data," *Computers & Graphics*, vol. 38, pp. 27–37, Feb. 2014.
- [27] Z. Y. Feng, J. J. Huang, W. H. Tang, and M. Shahidehpour, "Data mining for abnormal power consumption pattern detection based on local matrix reconstruction," *International Journal of Electrical Power & Energy Systems*, vol. 123, pp. 106315, Dec. 2020.
- [28] Y. Q. Gao, B. Foggo, and N. P. Yu, "A physically inspired data-driven model for electricity theft detection with smart meter data," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5076–5088, Sep. 2019.
- [29] X. Q. Min, Y. Huang, and Y. H. Sheng, "Automatic determination of clustering centers for "clustering by fast search and find of density peaks"," *Mathematical Problems in Engineering*, vol. 2020, pp. 4724150, Apr. 2020.
- [30] S. H. Yin, H. D. Yang, K. K. Xu, C. J. Zhu, S. Q. Zhang, and G. S. Liu, "Dynamic real-time abnormal energy consumption detection and energy efficiency optimization analysis considering uncertainty," *Applied Energy*, vol. 307, pp. 118314, Feb. 2022.
- [31] L. A. P. Júnior, C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. da Costa, and J. P. Papa, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, Nov. 2016.
- [32] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *International Journal of Electrical Power & Energy Systems*, vol. 101, pp. 301–310, Oct. 2018.
- [33] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri, "A tunable fraud detection system for advanced metering infrastructure using short-lived patterns," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 830–840, Jan. 2019.
- [34] C. Deng, K. Wu and B. Wang, "Residential Appliance Detection Using Attention-based Deep Convolutional Neural Network," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 2, pp. 621–633, Mar. 2022.

- [35] D. D. Zhang, B. A. Chen, H. Y. Zhu, H. H. Goh, Y. X. Dong, T. Wu, "Short-term wind power prediction based on two-layer decomposition and BiTCN-BiLSTM-attention model," *Energy*, vol.285, pp. 128672, Dec. 2023.
- [36] R. Sammouda and A. El-Zaart, "An optimized approach for prostate image segmentation using k-means clustering algorithm with elbow method," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 4553832, Nov. 2021.
- [37] J. T. Wu, Y. Yang, and J. S. Cheng, "A novel estimation method of friction coefficient for evaluating gear pitting fault," *Engineering Failure Analysis*, vol. 129, pp. 105715, Nov. 2021.
- [38] Y. Himeur, A. Alsalemi, F. Bensaali, and A. Amira, "Building power consumption datasets: survey, taxonomy and future directions," *Energy and Buildings*, vol. 227, pp. 110404, Nov. 2020.
- [39] Z. I. Bell, P. Deptula, E. A. Doucette, J. W. Curtis, and W. E. Dixon, "Simultaneous estimation of Euclidean distances to a stationary object's features and the Euclidean trajectory of a monocular camera," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4252–4258, Sep. 2021.
- [40] D. Simon and D. L. Simon, "Analytic confusion matrix bounds for fault detection and isolation using a sum-of-squared-residuals approach," *IEEE Transactions on Reliability*, vol. 59, no. 2, pp. 287–296, Jan. 2010.
- [41] S. C. Yi and L. Q. Yao, "A steady barycentric Lagrange interpolation method for the 2D higher-order time-fractional telegraph equation with nonlocal boundary condition with error analysis," *Numerical Methods for Partial Differential Equations*, vol. 35, no. 5, pp. 1694–1716, Sep. 2019.
- [42] X. Y. Ma and W. Shi, "AESMOTE: adversarial reinforcement learning with SMOTE for anomaly detection," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 943–956, Apr./Jun. 2021.



Dongdong Zhang received the B.Eng degrees in Electrical Engineering and Automation from the College of Electrical Engineering, Qilu University of Technology, Shandong, China, in 2013, received M.S. degree in Electric Power System and Automation from North China Electric Power University in 2016, and received Ph.D. degree in Electrical Engineering at the Xi'an Jiaotong University in 2019. In 2019–2020, he was a postdoctoral scholar with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and

Computer Engineering, University of Macau, Macao, China. Currently, he is a Professor with School of Renewable Energy, Inner Mongolia University of Technology. His research interests include the modeling and optimization of multiple energy system, energy market, and electrical machines and its driving system design.



Feng Jiang received the B.Eng degrees in Hefei University of Technology, Hefei, China, in 2019. He is currently pursuing the M.Eng. degree in Guangxi University, Nanning, Guangxi, China. His research interests include modeling of abnormal power consumption detection and artificial intelligence algorithm.



Hui Hwang Goh received the B.Eng (Hons.) and M.Eng. degrees in Electrical Engineering, and Ph.D. degree in Electrical Engineering from the University Teknologi Malaysia, Johor Bahru, Malaysia, in 1998, 2002, and 2007, respectively. He is currently a Professor of Electrical Engineering with the School of Electrical Engineering, Guangxi University, Nanning, China. His research interests include embedded power generation modeling and simulation, power quality studies, wavelet analysis, multicriteria decision-making, renewable energies, and dynamic equivalent. Dr. Goh is a Fellow of the Institution of Engineering and Technology, U.K. He is also a Fellow of the ASEAN Academy of Engineering and Technology and the Institution of Engineers, Malaysia, a Chartered Engineer under the Engineering Council United Kingdom, and a Professional Engineer under the Board of Engineers, Malaysia.



Tianhao Liu received the B.Eng and M.S. degrees in Electrical Engineering and Automation from the College of Electrical Engineering, North China Electric Power University in 2013 and 2016, respectively; and Ph.D. degree in Electrical Engineering from the Hong Kong University in 2020. After that, he worked as a scientist in State Grid in 2020 and in ABB Group Co., Ltd in 2021, respectively. Now, he works as a Senior Engineer in Huawei Digital Energy Technologies Co., Ltd. His research interests include motor system optimal operation, power

system simulation and economics, operation, and planning of multiple energy systems.



Hui Liu received his M.S. degree in 2004 and Ph.D. degree in 2007 from the College of Electrical Engineering at Guangxi University, China, both in Electrical Engineering. He worked in Tsinghua University as a postdoctoral fellow from 2011 to 2013 and in Jiangsu University as a faculty member from 2007 to 2016. He visited the Energy Systems Division at Argonne National Laboratory, Argonne, IL, USA, from 2014 to 2015. He joined the School of Electrical Engineering at Guangxi University in 2016, where he is a Professor and Deputy Dean. His

research interests include power system optimization, power system stability and control, electric vehicles, integrated energy systems, demand response, etc.



Thomas Wu received the Ph.D. degree in Electrical Engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1999. In 1999, he joined the University of Central Florida, Orlando, FL, USA, as an Assistant Professor, where he was promoted to an Associate Professor in 2005 and a Professor in 2011. He is currently a Professor of Electrical Engineering with the School of Electrical Engineering, Guangxi University, Nanning, China.